2022 8th International Conference on Systems and Informatics (ICSAI) | 978-1-6654-7448-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSAI57119.2022.10005470

Customer Portrait for Metrology Institutions Based on the Machine Learning Clustering Algorithm and the RFM Model

Xiaojing Zhang, Dongshuo Zhao, Yaran Li, Yudong Liu, Gang Hu
National Institute of Metrology, China
Beijing, 100029

***Abstract*—With the increasing intensity of competition in the current metrology testing market, building customer portrait is an effective way for metrology institutions to improve service levels to customers. This paper is based on the basic business data of a certain metrology institution. First, recency, frequency, monetary value model (RFM model), which is widely applied in customer relationship management, is improved. Further, it is combined with the business features of the metrology institution and used to build data feature engineering, which is closely related to the business data of the metrology institution and can reflect the data situation. Then, the data are analyzed through correlation test, standardized by Z-score, and clustered with three clustering algorithms, namely K-Means, DBSCAN, and AGNES, which are in SKLEARN database based on Python. After that, the clustering results are compared. In the clustering process, the elbow method and method for traversing the silhouette coefficient are used to determine the optimal value of the clustering algorithm. Finally, with the analysis of clustering results, the customers' features of the metrology institution are signed and the customer portrait is built, which provides data analysis methods, tools and decision basis for the metrology institution to offer better services.**

***Keywords*—machine learning; clustering algorithm; customer portrait; RFM model.**

I. Introduction

Customers are the major competitive resources for metrology institutions in the industrial metrology market, moreover, high-quality customers are directly related to the business growth of metrology institutions [1]. On the one hand, the competition in the industrial metrology market is currently becoming increasingly fierce. Therefore, metrology institutions attempt to maintain customer satisfaction and obtain new high-quality customers to achieve business growth by adopting multiple approaches, such as improved testing services, low testing prices and high-quality certificates. However, the business level of metrology institutions is limited by service resources, such as the number and level of customer service specialists and the delivery capacity of testing instruments. Maximizing existing service resources and identifying the service direction to improve the competitiveness in the industrial metrology market and finally achieve a win-win situation with a relatively balanced investment in an expanded testing market and service resources is clear. Competition in the industrial metrology market has been a critical factor for metrology institutions [2]. Combined with the industrial features of metrology, this paper improves the traditional recency, frequency, monetary value (RFM) model, and reselects the customer eigenvalues suitable for the metrology industry to endow the model with characteristics close to the metrology industry. Afterward, the data are clustered by three clustering algorithms, namely K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and AGglomerative NESting (AGNES), which are learned by Python in SKLearn database, to determin the relationship among the characteristics of customer business data. Metrology institutions then build customer portrait[1] based on customer types after rating and dividing them into different types according to the clustering results, and utilizing existing resources to formulate service strategies to facilitate the transformation into serving high-quality customers.

II. Feature Engineering

Feature engineering [3] is a process of transforming raw data into data with a superior expression of characteristics, which usually includes data preprocessing, feature selection, and dimensionality reduction. Applying features to models through feature engineering can not only reflect issues represented by the data itself but also improve the model calculation accuracy of invisible data. Feature engineering accounts for the largest and the most important part of machine learning. Therefore, the selection of data features has a direct impact on the construction of models and final analysis results of data.

A. RFM Model

As an important manifestation and tool for measuring the value and value creation ability of customers, the RFM model selects the indicators of consumer behavior to determin and evaluate their values [4] [5]. The model describes the overall value of the customer through the following three indicators: the recent consumer behavior (Recency), the overall consumption frequency (frequency) and total consumption (monetary) [6] [7]. In the metrology industry, the three indicators of the RFM model can be understood as follows: recency is the

[1] Customer portrait refers to the following: in the process of market consumption, customers accumulate various types of data through consumption and other behaviors, such as customer information, consumption information, and behavior data. These data carry most of the customer's information, which has considerable value for the seller to analyze the customer data. The application of big data to establish the customer portrait to study the customer attributes and labels and then clarify the sales means is gradually becoming the key to market competition.

978-1-6654-7448-1/22/$31.00 ©2022 IEEE

TABLE I
CORRELATION COEFFICIENT BETWEEN DATA EATURES

	Feature B	Feature C	Feature F	Feature G
Feature B	1	0.643309	0.085753	0.0414759
Feature C	0.643309	1	0.063055	0.00022
Feature F	0.085753	0.063055	1	0.05951
Feature G	0.014759	0.00022	0.05951	1

TABLE II
CORRELATION COEFFICIENT BETWEEN DATA FEATURES

Items	Quality	Minimum	Maximum
Z-score (Feature B)	970	-0.44356	11.74395
Z-score (Feature F)	970	-0.96681	2.864133
Z-score (Feature G)	970	-0.24493	17.21185

last time that the customer tests the instrument; frequency is the testing frequency of customers; monetary is the sum of the testing expenses spent by the customer in the metering institution.

B. Improvement of the RFM Model

The data features in this paper including the following: Feature A: full name of the customer Feature B: total annual test cost Feature C: total annual quantity of test bench Feature D: last time that the customer tests the instrument Feature E: testing frequency of the customer Feature F: geographical distance between customers and metrology institutions Feature G: growth of annual test cost compared with that of last year Compared with ordinary consumption market behaviors, the metrology industry has cyclical characteristics. Various factors such as the wear and tear of aging of a measuring instrument while use, as well as the environment, temperature, and humidity during use, will influence its data and relatively cause deviations. Therefore, errors must be corrected through periodic detection according to the cycles of different measuring instruments, to ensure the accuracy and reliability of the measuring results. This feature, indicates that demonstrating the characteristics of customer business data detection for the last testing time of the customer in Feature D and the testing frequency of the customer in Feature E is impossible. Therefore, the two features are not considered in this paper.

C. Correlation Test

Correlation tests on the characteristics of Features B, C, F, and G are further conducted. Generally, a correlation coefficient that is larger than 0.7 is considered a strong correlation. Meanwhile, anything between 0.5 and 0.7 is a moderate correlation, and anything less than 0.4 is considered a weak correlation. The pairwise correlation coefficients of the four eigenvalues are shown in Table 1. The table reveals that the correlation coefficient between the total annual test of Feature B and the total annual quantity of the test bench of Feature C is 0.64, which has a linear correlation. Feature B is used in this paper to perform the follow-up analysis.

Finally, the improved RFM model is constructed as follows: the growth of the total annual test in Feature B, the geographical distance between customers and the metrology institution in Feature F, and the growth of annual test cost compared with that of last year in Feature G compared with last year.

D. Normalization

Normalization means to proportionally scale the data, transform them into a dimensionless pure numerical value, and place them into a small specific interval. Data features have different dimensions and orders of magnitude; thus, if substantial differences are found among the levels of each indicator and the original index is directly used for analysis, then the impact of indicators with a high numerical value in the comprehensive analysis will be highlighted, while that of indicators with relatively low numerical value will be weakened. Therefore, normalization is often used in data analysis to ensure the reliability of results. The original data are standardized through typical methods, such as range standardization and Z-score standardization. Z-score standardization is adopted in this study, and its formula is expressed as follows:

$$Z = \frac{(x-\mu)}{\sigma}, \tag{1}$$

where μ is the average number, and σ is the standard deviation of the data group. The core statement of SKLearn is as follows:

```
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler().fit(X)
```

The descriptive statistics of the data after being Z-score standardized are listed in Table 2.

III. CLUSTERING ALGORITHM

The clustering algorithm belongs to unsupervised learning in machine learning. The purpose of this algorithm in supervised learning is to determine the classification of samples according to their characteristics considering clear classification objectives. Therefore, supervised learning must be defined and classified by humans. Unsupervised learning aims to determine the structure of data sets without advanced classification and cluster them into groups automatically. A number of clustering algorithms are provided in the machine learning SKLEARN database. Three representative algorithms, namely K-Means, DBSCAN, and AGNES, are adopted in this paper for clustering.

A. K-Means Clustering Algorithm

1) Clustering Principles: The K-Means clustering algorithm is a clustering analysis of iterative solutions based on partition clustering [8] [9] [10]. This algorithm contains the following steps: specifying k objects as the initial cluster center; calculating the distance between each object and the cluster center and selecting the center with the shortest distance; classifying the sample points into the cluster center represented by the selected center. The cluster centers and the assigned objects represent a cluster, each time being allocated a sample, and the cluster center is recalculated in accordance with the existing objects in the cluster and form a new cluster. This process is repeated until the maximum number

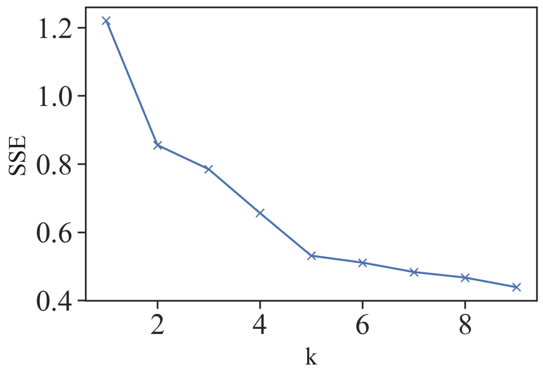

Fig. 1. The relationship between k value and the sum of the squared of errors

of iterations is reached or the updated cluster center is almost consistent with the original cluster center and forms a fixed point.

2) Model Evaluation: K-Means has many advantages [11][12]: for example, its algorithm principle is easy to understand and has a considerable clustering effect on large volumes of scale dates owing to its fast calculation speed. However, the limitation of this algorithm lies in its strong subjectivity owing to the k value; thus, the cluster center must be artificially specified. The elbow method or silhouette coefficient is usually adopted to determine k values to simultaneously obtain an improved clustering effect and evaluate the optimal k value.

a) : Elbow Method The division of observation points will be refined and the aggregation degree of each category will gradually rise with the steadily increasing k value of the cluster center. Therefore, the sum of the squared errors will gradually decrease. The formula of the sum of the squared error is as follows:

$$SSE = \sum_{i=1}^{k}\sum_{p\in C_i}|p-m_i|^{2}, \tag{2}$$

where Ci is the first i cluster, p is an observation point in this cluster, and mi is the cluster center. The sum of the squared errors represents the clustering effect. When k is less than the real cluster number, the degree of polymerization of each category increases, and the sum of the squared errors then decreases with the increase in k value. When the k value is close to optimal clustering, the increasing k value will quickly reduce the clustering effect, and the decreasing range of the sum of the squared errors will stabilize. The sum of the squared errors and k value approximate the elbow shape, and the k value in the elbow is that which is close to the optimal clustering effect. This model uses the elbow method for determining the k value, which is shown in Fig. 1.

Fig. 1 shows that the k value can be taken as 2 or 5.

b) : Silhouette Coefficient The silhouette coefficient is another method of evaluating the clustering effect whose value ranges from -1 to 1. A value close to 1 indicates an improved clustering effect. Combining the two factors of cohesion and resolution, silhouette coefficient can be used to evaluate the impact of different algorithms or operation modes of an algorithm on clustering effects based on the same original data. The clustering algorithm divided the data set to be clustered into k clusters [13][14][15]. Each vector in the cluster has a silhouette coefficient, which should be calculated. For one of the point i:

TABLE III
The relationship between k values and silhouette coefficient

k	Silhouette Coefficient
2	0.536889485
3	0.564795679
4	0.618839663
5	0.728295512
6	0.681259916
7	0.687647106
8	0.69090163

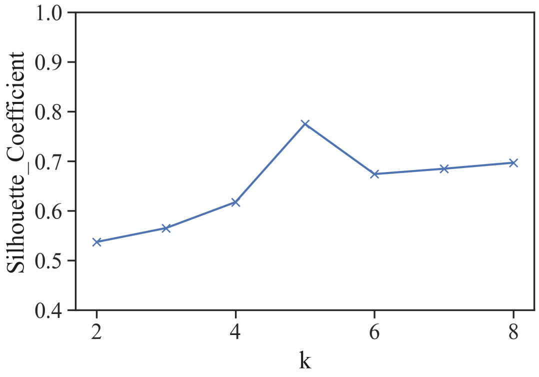

Fig. 2. The relationship between k values and silhouette coefficient

$$a(i) = average \tag{3}$$

(The distance from vector i to all other points in the cluster to which it belongs)

$$b(i) = min \tag{4}$$

(The average distance from vector i to all points in the cluster to which it does not belong)
Then, the silhouette coefficient of vector i is as follows:

$$S(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \tag{5}$$

This paper calculated the silhouette coefficients of k values from 2 to 8 through k value traversal to evaluate the optimal k value. The experimental results are shown in Table 3 and Fig. 2.

3) Visualization of the Optimal Clustering Effect: Five K-Means clustering effects are displayed after the optimal k value is taken. The visualization of clustering effects is shown in Fig. 3.

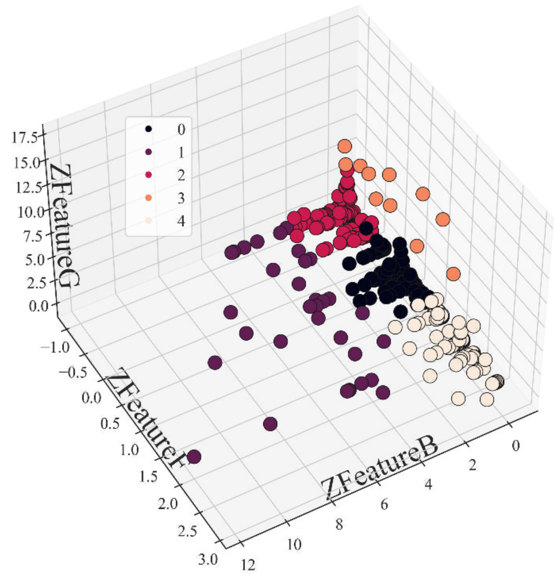

Fig. 3. Visualization of the optimal K-Means clustering effect

B. DBSCAN Clustering Algorithm

1) Algorithm Principle: The DBSCAN clustering algorithm is a data clustering algorithm based on density. The scan radius eps and the minimum number of included points MinPts should be set first in the algorithm. From any observation point, the number of objects (points or objects in other spaces) contained in a certain area (within eps) of the cluster space should be no less than a given threshold (MinPts). Then, whether this point confirms to the core point is observed under this condition. If it is the core point, then it should be recorded as the core point and all observation points contained by this point should be identified; if this point is not the core point, then this point is possibly the observation point within the scanning radius of other core points or serves as a noisy point. Afterward, whether nonobserved points are core points is continuously determined, and the above process is repeated. This process was continued until all observation points are marked as core points, scanned observation points, or noisy points.

The DBSCN clustering algorithm can also be used to find clusters in any shape in spatial data sets with noise without setting the number of clusters. However, this algorithm also has some disadvantages. Finding clusters in different densities is difficult, and the two parameters, namely eps and MinPts, must be set artificially.

2) Value Verification and Model Evaluation: Simultaneously, the silhouette coefficient should be optimal and the proportion of noisy points in the total observation points should be the smallest to evaluate the two optimal parameters of eps and MinPts. From 0.1 to 11, the eps value is traversed in 0.1 steps; from 2 to 10, the MinPts value is traversed in 1 step. The experimental result shows a total of 584 results with a silhouette coefficient larger than 0.5. The class 1 or 2 cluster is not adopted considering the absence of significant effects on the actual work. Table 4 shows 19 clusters whose effects are larger than 2.

TABLE IV
CORRESPONDENCE BETWEEN SILHOUETTE COEFFICIENTS AND VALUES OF EPS AND MINPTS TRAVERSE

eps	MinPts	Sihouette Coefficient	Noise_Ratio	n_clusters
1.8	2	0.800142821	0.206%	3
1.6	2	0.799240531	0.4124%	3
1.7	2	0.799240531	0.4124%	3
1.3	2	0.732787329	0.5155%	5
1.5	2	0.732755261	0.4124%	4
1.1	3	0.730284457	1.2371%	3
1.2	3	0.730284457	1.2371%	3
1.3	3	0.729491985	0.9278%	3
1.4	2	0.725026878	0.9278%	5
1.4	3	0.721909822	0.8247%	3
0.7	3	0.666969314	2.3711%	3
0.8	5	0.665268617	2.8866%	3
0.7	4	0.645920101	2.3711%	4
1.2	2	0.642800933	0.8247%	5
0.9	2	0.641430231	1.2371%	4
1	2	0.641227494	1.1340%	4
1.1	2	0.641050396	1.0309%	4
0.7	5	0.621958473	3.2990%	4
0.7	6	0.619132536	3.9175%	3

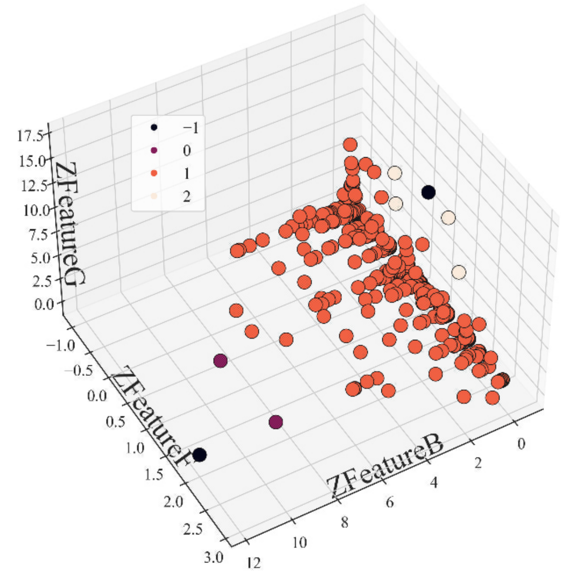

Fig. 4. Visualization of the optimal clustering effect of DBSCAN

3) Visualization of Optimal Clustering Effect: The clustering effect of DBSCAN is class 3 (-1 is the noisy point) after the optimal valuing of the two parameters, eps and MinPts, as shown in Fig. 4.

Fig. 4 shows that the number of clusters marked in 1 is far larger than the two other clusters. The clustering purposes of using the two characteristics of Features F and G to maximize the homogeneity of the research objects and the heterogeneity of the objects between clusters are not achieved. Therefore, finding clusters in different densities is difficult for the DBSCAN clustering algorithm due to its algorithm features. Thus, this algorithm is not applied to the data feature

TABLE V
THE RELATIONSHIP BETWEEN K VALUES AND SILHOUETTE COEFFICIENT

k	Silhouette Coefficient
2	0.516894473
3	0.583544732
4	0.679051911
5	0.707032347
6	0.773298553
7	0.713836808
8	0.71075396

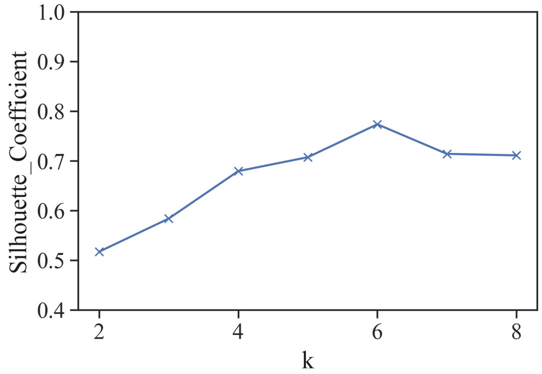

Fig. 5. The correspondence between the k value and the silhouette coefficient

model constructed in this paper.

C. AGNES Clustering Algorithm

1) Algorithm Principle: AGNES clustering algorithm is a representative hierarchical clustering method, which adopts the algorithm of the bottom-up aggregation strategy [16]. The algorithm mainly aims to take every object as an initial cluster initially and calculate the distance between any two clusters to find the two clusters with the shortest distance and combine them to form a new cluster. These clusters are then combined step by step by principles, and this process is repeated until all objects are satisfied with the number of clusters. The conditions of process termination are as follows:

A. Design a minimum distance threshold; if the shortest distance is larger than the threshold, then combining clusters is no longer necessary.

B. Restrict the number of clusters to k; terminate the clustering once the number of clusters reaches k.

2) Value Verification and Model Evaluation: A k value is also artificially set in AGNES; therefore, traversing the silhouette coefficient is adopted again to set the k value. The correspondence between the k value and the silhouette coefficient is shown in Table 5 and Fig. 5.

3) Visualization of the Optimal Clustering Effect: Table 5 and Fig. 5 show six clustering effects after the optimal value. The visualized effect is shown in Fig. 6.

D. Comparison of Clustering Effects

According to all the above analyses, in these three machine learning clustering algorithms, namely K-Means, DBSCAN, and AGNES, the DBSCAN clustering algorithm is unsuitable for the model constructed in this paper, and its clustering effects do not achieve the clustering purposes. The K-Means and ANGES clustering algorithms have achieved remarkable clustering effects. AGNES has a higher silhouette coefficient than K-Means. However, the comparison between Figs. 3 and 6 reveal that the differences between clusters are evident when using the K-Means clustering algorithm. As a result, the K-Means clustering algorithm is most suitable for the data of the model constructed in this paper.

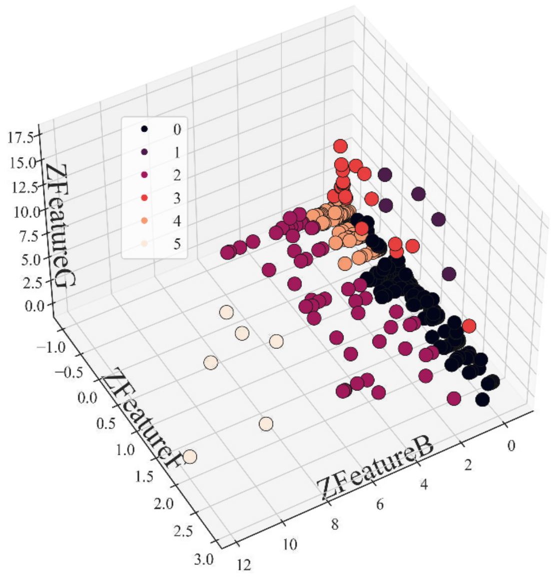

Fig. 6. Visualization of the optimal clustering effect of AGNES

Type	Number of Customers	Total Test Cost	Physical Distance	Growth of Test Cost	Service Strategy
0	137	low	far	low	maintain
1	492	low	close	low	retain
2	11	low	average	high	developing
3	33	high	average	low	high-quality
4	297	low	medium	low	retain

IV. PORTRAIT OF CUSTOMERS IN METROLOGY INSTITUTIONS

Based on the K-Means clustering algorithm effects in Fig 3, the differences among customers are reflected, and the customer portrait for these metrology institutions is obtained, shown in Table 6 and Fig. 7 [17].

After clustering, the differences of all types of customers are close to the situation of the metrology industry, that is, the feature that few high-quality customers create most business income. In the portrait of customers, the total test cost of type 3 customers is far higher than that of other types. Therefore, this type of customer should be the most valuable and closely business-connected customer of the metrology institutions [18][19]. With low total test cost and the high growth of test cost, customers in type 2 are in the initial stage of cooperation with the metrology institution. Therefore, the metrology institution should strengthen cooperation and

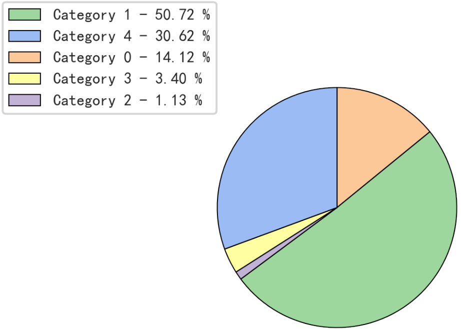

Fig. 7. Occupation of the quantity of all types of customers

communication in combination with the customer scale and their reputation. Some customers have considerable business potential. Combined with several service strategies, such as instrument transportation, pick-up, and delivery, service resources can be invested in these customers. The quantity of customers in types 0, 1, and 4 occupies over 80 percent of the total quantity of customers in this metrology institution. This similarity can be further understood as follows: customers with long physical distances are the standard regular traceability of the metrology industry system, and the sum of their testing expenses is maintained on average throughout the year. Other customers have small business scales or cooperation with other metrology institutions. Therefore, realizing business growth is difficult for the metrology institution despite investing in service sources. Consequently, business service sources should not be overinvested but should be combined with the aforementioned finding and maintain the present service strategy. According to the above analysis of the portrait of customers, in the current environment of increasingly fierce competition in the metrology industry and the normalization of the COVID-19, this metrology institution should take the initiative to change the business service strategy for customers [20], and concentrate human, material and other service resources to serve the minority high-quality customers.

V. Conclusion

This paper reconstructed a RFM business data analysis model which is more closed to the metrology industry features through considering the geographical distance between customers and metrology institution and the growth of annual test cost. The data are clustered by machine learning clustering algorithms. After taking the correlation test and standardization to the data, the clustering effect of the K-Means clustering algorithm is considered optimal, and the portrait of customers is built according to the effect, which provides data analysis tools and a decision-making basis for metrology institutions to improve their service competitiveness in the testing market.

Acknowledgment

This research is funded by the"Research on Detection Business Data Analysis and Customer Profiling" (APT2201-13).

References

[1] R. Xue, X. Su, Z. Tu, "Research on Hierarchical Service Model of Social Customers in Metrology Institutions," Manage. Admin., vol. 10, pp. 18-22, 2019. (references)

[2] X. Zhang, Q. Sun, C. Lv, "Brief Talk of the Current Situation and Development of Domestic Metrology," Institut. Indus. Measu.r, vol. 6, pp. 16-17, 2016. (references)

[3] Y. Tang, S. Sun, "Feature Engineering Method in Machine Learning," Aut. Appl. Technol., vol. 12, pp. 3, 2020. (references)

[4] Y. Chen, L. Shi, X. Zhang, Z. Xie, "Research on Evaluation System and Influencing Factors of Popular Books in University Library Based on Modified RFM Model," Res. Lib. Sci., vol. 10, pp. 58-68, 2020. (references)

[5] H. Li, S. Wang, "Modeling and Verification of Mall Member Portrait Based on RFM Mode," J. Zhengzhou. Railway. Vocat. Tech. Coll., vol. 31, pp. 14-24, 2019. (references)

[6] Z. Yi, X. Hao, "Customer stratification theory and value evaluation—analysis based on improved RFM model," J. Intel. Fuzzy. Syst., vol. 40, pp. 4155-4167, 2021. (references)

[7] G. Sun, X. Xie, J. Zeng, W. Jiang, "Using improved RFM model to classify consumer in big data environment," Inter. J. Embed. Syst., vol. 14, pp. 54, 2021. (references)

[8] C. Tian, W. Yang, D. Yang, Y. Wang, "Analysis and Research on Students' Behavior Based on Comprehensive Data of Colleges and Universities Under the Background of K-means and DBSCAN Clustering," Algor. Sci. Tech. Innov., vol. 32, pp. 86-88, 2020. (references)

[9] P. Ma, "Research on Data Mining and Customer Segmentation Based on K-means Algorithm," Stat. Longitud. Latitud., vol. 11, pp. 66-67, 2019. (references)

[10] F. Zhu, Y. Wang, "Research on User Information Clustering and Prediction Based on K-means and Neural Network Machine Learning Algorithm," Inform. Sci., vol. 7, pp. 83-90, 2021. (references)

[11] K. Zhou, s. Yang, "Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering," Pattern. Anal. Appl., vol. 23, pp. 455-466, 2020. (references)

[12] H. Yu, G. Wen, J. Gan, W. Zheng, "Self-paced Learning for K -means Clustering Algorithm," Pattern. Recogn. Lett., vol. 132, pp. 69-75, 2020. (references)

[13] X. Shao, Y. Jia, W. Zhang, J. Ding, J, "Data analysis based on K-Means clustering algorithm," Sci. Techn. Innov., vol. 23, pp. 85-86, 2021. (references)

[14] F. Bu, Z. Chen, Q. Zhang, L. Yang, "Incomplete high-dimensional data imputation algorithm using feature selection and clustering analysis on cloud," J. Supercomput. vol. 72, pp. 2977-2990, 2016. (references)

[15] D. Li, "Research on Renewal Intention of Auto Insurance Customers Based on Machine Learning Algorithm," CAAI. Trans. Intell. Techn., vol. 10, pp.211-213. Inst. Eng. Techn., 2021. (references)

[16] L. Wilmer, E. J. Llanos, R. Guillermo, C. F. "Suárez, How frequently do clusters occur in hierarchical clustering analysis? A graph theoretical approach to studying ties in proximity," J. Cheminformatics., vol. 8, pp. 4, 2016. (references)

[17] K. Sanjib, K. Sanjit, M. Sasmita, "Improvement of CRM using Data Mining: A Case study at Corporate Telecom Sector," Int. J. Comput. Appl., vol. 178, pp. 12-20, 2019. (references)

[18] D. Zhao, Y. Liu, X. Zhang, T. Y. Liu, "The Business Structure of a Metrology Institution Based on Linear Regression Analysis," ICSAI. 2021 - 7th. Inter. Conf. Systems. Informatics. 2021. (references)

[19] Y. Liu, X. Zhang, D. Zhao, Y. Pu, "Definition and Segmentation of Key Clients in the N Metering Institution Based on Data Analysis," Lect. Notes. Data. Eng. Commun. Techn. vol. 89, pp. 471-485, 2022. (references)

[20] Y. Liu, Z. Hu T. Liu, D. Zhao, "Discussion on improving the receiving efficiency and quality of measuring instruments. China Metrology," vol. 89, pp. 471-485, 2022. (references)